Structure, Vol. 10, 1613–1618, December, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00894-8

Application of NMR in Structural Proteomics: Screening for Proteins Amenable to Structural Analysis

Ways & Means

Till Rehm, Robert Huber, and Tad A. Holak[1]
Max-Planck-Institut für Biochemie
Am Klopferspitz 18a
D-82152 München
Germany

Summary

In the time of structural proteomics when protein structures are targeted on a genome-wide scale, the detection of "well-behaved" proteins that would yield good quality NMR spectra or X-ray images is the key to high-throughput structure determination. Already, simple one-dimensional proton NMR spectra provide enough information for assessing the folding properties of proteins. Heteronuclear two-dimensional spectra are routinely used for screenings that reveal structural, as well as binding, properties of proteins. NMR can thus provide important information for optimizing conditions for protein constructs that are amenable to structural studies.

Introduction

It has been widely assumed that nuclear magnetic resonance spectroscopy will play an important role in structural proteomics, complementing X-ray crystallography for small- and medium-size proteins (below 30 kDa) [1, 2]. About 17% of the structures deposited in the Protein Data Bank, most of which do not have corresponding crystal structures, have been solved by NMR spectroscopy [2, 3]. We believe, however, that NMR will remain a "poor daughter" of the X-ray method in determining structures of proteins. Nevertheless, NMR can deliver strong results in several areas of structural biochemistry. It is the basis for a wide range of experiments to determine structure-function relationships [4], to find binding partners with their specific binding sites [5], to investigate dynamics of proteins [6], to distinguish multiple conformations [7], to compare apo and holo forms of proteins and map the binding sites of their cofactors [8], or to determine pKa values of ionizable groups [9], to name just a few. A series of spectra taken under different conditions may be used to monitor aggregation and formation of amyloid fibrils [10], to determine K_D values of binding partners [4], or to track hydrogen exchange with real-time NMR in proteins dissolved in D_2O [11]. The ability to detect ligands binding only very weakly to target molecules has made NMR also increasingly important in drug discovery [12, 13].

In this report we focus on the application of NMR for screening for protein samples that are suitable for structure elucidation by both NMR spectroscopy and X-ray crystallography. Securing "well-behaved" samples is expected to be the rate-determining step in any structural proteomics project [14, 15].

In a recent investigation of roughly 500 proteins from the genome of a single organism, Christendat et al. [14, 16] found that only ~10%–15% of these proteins yielded samples that were of sufficient quality for structural analysis by either NMR spectroscopy or X-ray crystallography. Clearly, a method to screen for these well-behaved proteins as well as to optimize the samples of the many others is urgently needed.

The unique strength of NMR lies in its capability to semiquantitatively estimate unstructured regions of the polypeptide chain in the otherwise partially folded protein and to identify proteins that are heterogeneous because of aggregation or other conformational effects. We will illustrate the various applications of NMR in structural proteomics using examples from our own work; this article is not intended to review the literature, but rather to provide typical examples, of these NMR applications.

One-Dimensional NMR

A simple one-dimensional proton experiment, the most basic spectrum in NMR spectroscopy that can be acquired in a short time (usually not longer than a few minutes) for samples as dilute as 0.01 mM, already contains a great amount of information. The lower panel of Figure 1 shows an example of an unfolded protein with a large and broad signal at approximately 8.3 ppm. An unfolded protein shows a small dispersion of the amide backbone chemical shifts [17]. Particularly, the appearance of intensities at chemical shifts near ~8.3 ppm is an excellent indicator for a disordered protein, as this is a region characteristic of backbone amides in random-coil configuration. On the other hand signal dispersion beyond 8.5 ppm (8.5–11 ppm) proves a protein to be folded. Because of the different chemical environment and, thus, the varying shielding effects, the resonances of the single protons will be distributed over a wide range of frequencies. A typical intensity pattern of a folded protein is shown in the upper panel of Figure 1. Following the same argument, in the aliphatic region of the spectrum, between 1.0 and −1.0 ppm, a large signal dispersion versus a steep flank of the dominant peaks at approximately 1 ppm separates a structured protein from an unfolded protein (Figure 1, upper and lower panels, respectively).

Close inspection of one-dimensional spectra will also yield quantitative information on the extent of folding in partially structured proteins or their domains. Figure 2 shows two spectra of a 20 kDa protein. In the upper spectrum a mixture of approximately 50%-folded and 50%-unfolded protein can be identified by observing both the signal dispersion and the prominent peak at 8.3 ppm. The lower spectrum shows the same sample after removal of the unfolded macromolecules by gel

[1]Correspondence: holak@biochem.mpg.de

Key words: NMR; X-ray crystallography; screening; protein characterization; proteomics; structural analysis

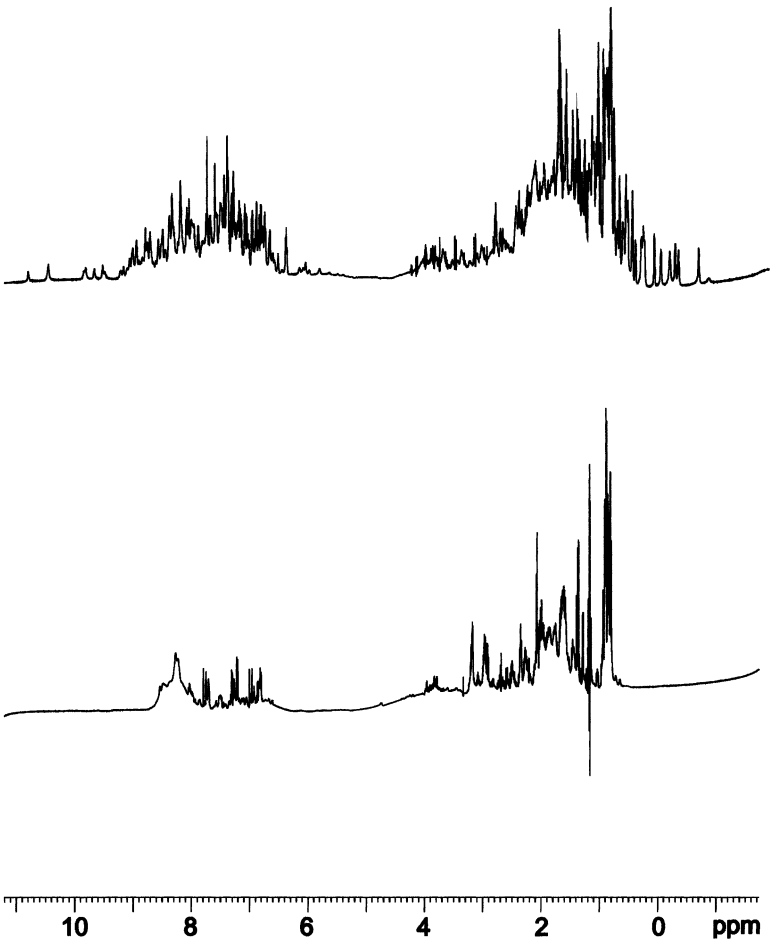

Figure 1. Characterization of Protein Structures by One-Dimensional NMR Spectroscopy

(A) A typical one-dimensional proton NMR spectrum of a folded protein with signal dispersion downfield (left) of 8.5 ppm and upfield (to the right) of 1 ppm. Spectra show the N-terminal 176-residue domain of the cyclase-associated protein (CAP) at pH 7.3.

(B) An unfolded protein sample. Strong signals appear around 8.3 ppm, the region characteristic for amide groups in random-coil conformation. No signal dispersion is visible below approximately 8.5 ppm. Also, to the right of the strong methyl peak at 0.8 ppm, no further signals show up. The sample is an unfolded domain of the IGF binding protein 4 (IGFBP-4, residues 147–229).

filtration. The "random-coil peak" disappeared, and the signal pattern is that of a completely structured protein.

While the signal dispersion of the resonances is generally connected to folding, aggregation can be detected by observing the line width of the signals. Because of faster relaxation mechanisms, the NMR signal from larger molecules will decay much faster than that from smaller ones [18]. This, in turn, will produce broader

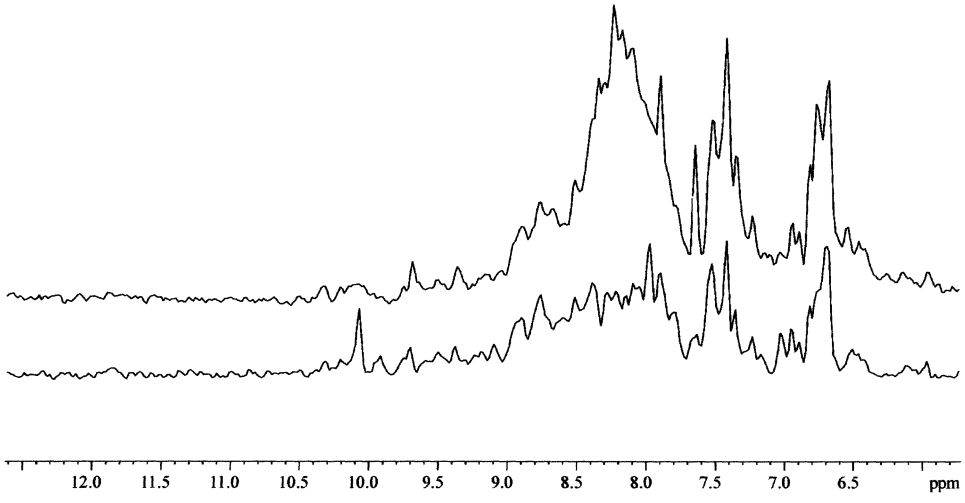

Figure 2. The Amide Region of a 20 kDa Protein

The upper trace shows a 1:1 mixture of folded and unfolded proteins; the lower trace shows the same sample after removal of the unfolded proteins by gel filtration.

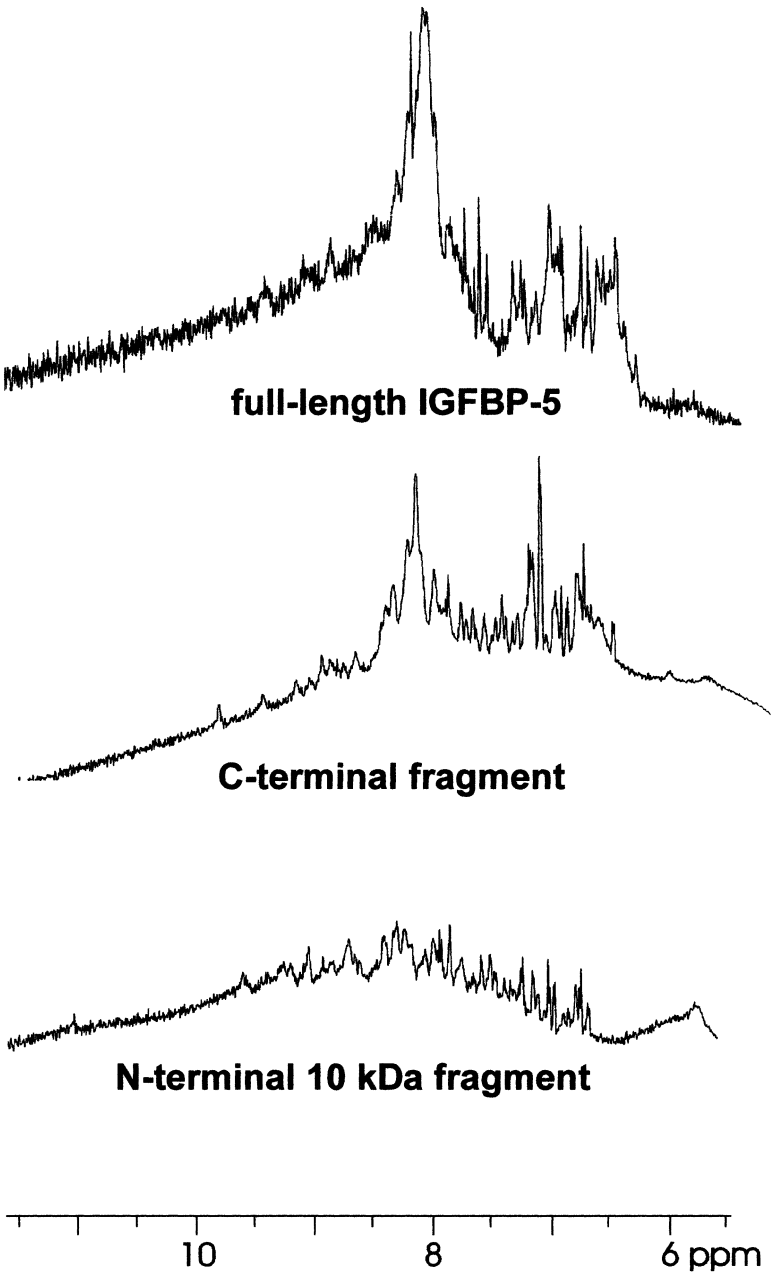

Figure 3. Amide Region of One-Dimensional Spectra of the IGF Binding Protein-5

The upper, middle, and lower panels show the full-length protein of 246 amino acid residues, a C-terminal fragment of 112 residues, and an N-terminal fragment of 94 residues, respectively.

lines for the resonances of larger molecules. Thus, the line widths of the signals in any NMR spectrum are correlated to the size of the molecule. Both these aspects may be appreciated in Figure 3. The upper panel shows the spectrum of the 246-residue IGFBP-5 [19] that exhibits a rather large peak at the random-coil value of 8.3 ppm and some signals downfield (that are shifted to higher parts per million values), close to the noise level. The IGFBP-5 protein comprises conserved N- and C-terminal domains of 90 and 112 amino acids, respectively, and a central domain of 40 amino acids. Spectra of the C-and N-terminal fragments of the same protein (Figure 3, middle and lower panels, respectively) show that there is an unstructured region located in the C-terminal fragment. The N-terminal fragment shows nice signal dispersion, characteristic of a structured protein. Again, in this example, the quantitative information on the extent of folding that is available from 1D NMR can be appreciated. The full-length protein is only about 50%–60% folded. This spectrum (Figure 3, upper panel) may be seen as a superposition of the two other spectra, which show the C-terminal fragment, which is about 30% folded, and the fully folded N-terminal fragment (middle and lower panels, respectively; the rest of the unstructured residues originate from the central domain of IGFBP-5).

Note that the line width of the individual signals has also improved dramatically in the smaller fragments. With the line width from known monomeric proteins of a given size as a reference, the observation of the line

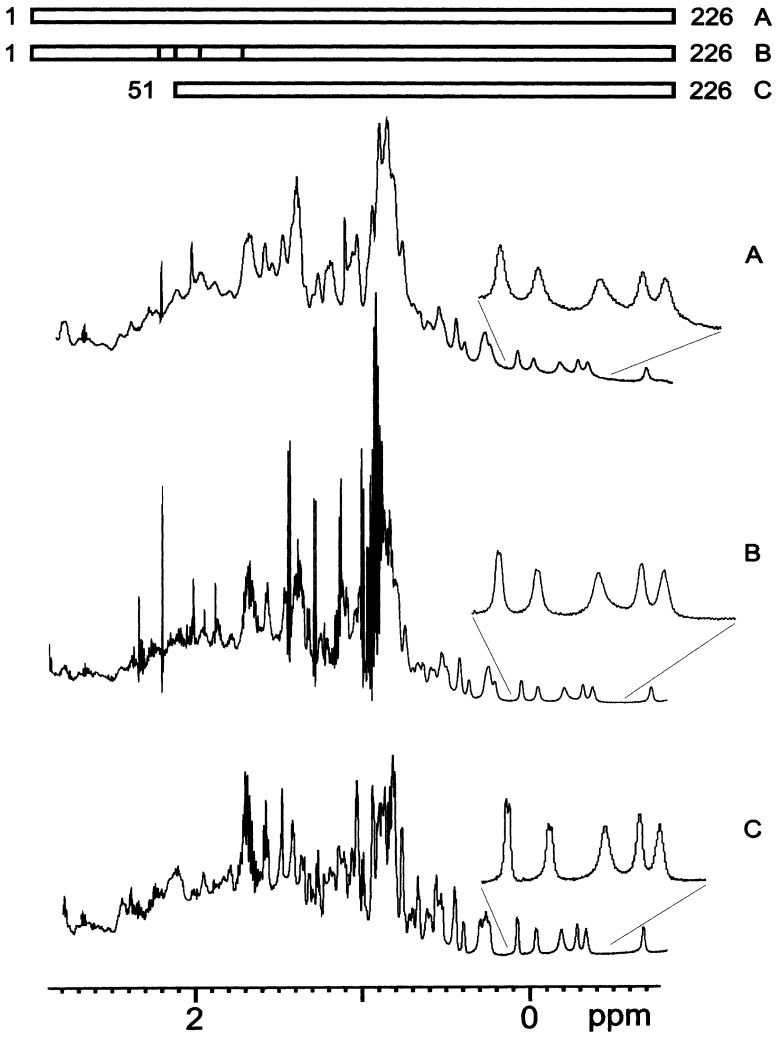

Figure 4. The Aliphatic Region of One-Dimensional Spectra of the Cyclase-Associated Protein (CAP)

(A) The N-terminal 226-residue construct, as indicated in the scheme above.

(B) A mixture of several constructs of different length (residues 1–226, 44–226, 51–226, and 56–226). The overlap leads to broader lines than in spectrum (A). Short peptides give rise to sharp signals around 1 ppm.

(C) The stable core of the protein (residues 51–226) only. The sharp signals from impurities are removed, and the line width is substantially improved compared with spectrum (A).

width in a one-dimensional spectrum will also yield information on the molecular weight and aggregation of the molecule under investigation. Furthermore, attempts to prevent aggregation by, for example, dilution of the sample, addition of mild detergents, such as CHAPS, or lowering the pH value can thus be monitored by NMR to find optimal sample conditions [19–21]. While the extent of folding is crucial both for X-ray crystallography and NMR, aggregation is not. Actually, some proteins that yield rather poor NMR spectra because of aggregation or low solubility might give excellent crystals, as did $p19^{INK4d}$ [22, 23]. Thus, sample conditions that are optimal for crystallography might not necessarily be optimal for NMR spectroscopy and vice versa. This fact does not, however, reduce the value of insights given by NMR for crystallography. In order to assess whether conditions that are optimal for NMR are also optimal for crystallography, one would have to sample a large number of proteins. We were informed that such large-scale comparisons are under way in the Northeast Structural Genomics Conosortium in the United States and at the RIKEN structural genomics project in Japan.

In comparing Figures 2 and 3, we have to point out that distinguishing between a protein that is only partially folded and a mixture of folded and unfolded proteins is difficult with NMR, without having additional information from, for example, gel filtration or other biochemical methods.

One-dimensional spectra may additionally provide information on α-helical or β strand structures in a protein. The C^{α} protons in a helix display few resonances in the region between 5 and 6 ppm, while those in a β sheet resonate in this region [24].

The use of one-dimensional spectra to screen for optimal, fully folded protein fragments may be illustrated by the example of the cyclase-associated protein (CAP) [25]. The initial construct of the protein, comprising 226 amino acids, showed considerable line width and would not crystallize (Figure 4A). After the sample was left at room temperature for 7 days, another one-dimensional spectrum showed, on the one hand, degraded peptide fragments but, on the other hand, still showed broad lines, not in agreement with the expected shorter protein fragment (Figure 4B). Mass spectrometry revealed the presence of several protein fragments of different lengths, ranging from 226 to 173 amino acids. Thus, the very sharp peaks around 1 ppm could be attributed to the cleaved peptide fragments, while the line width

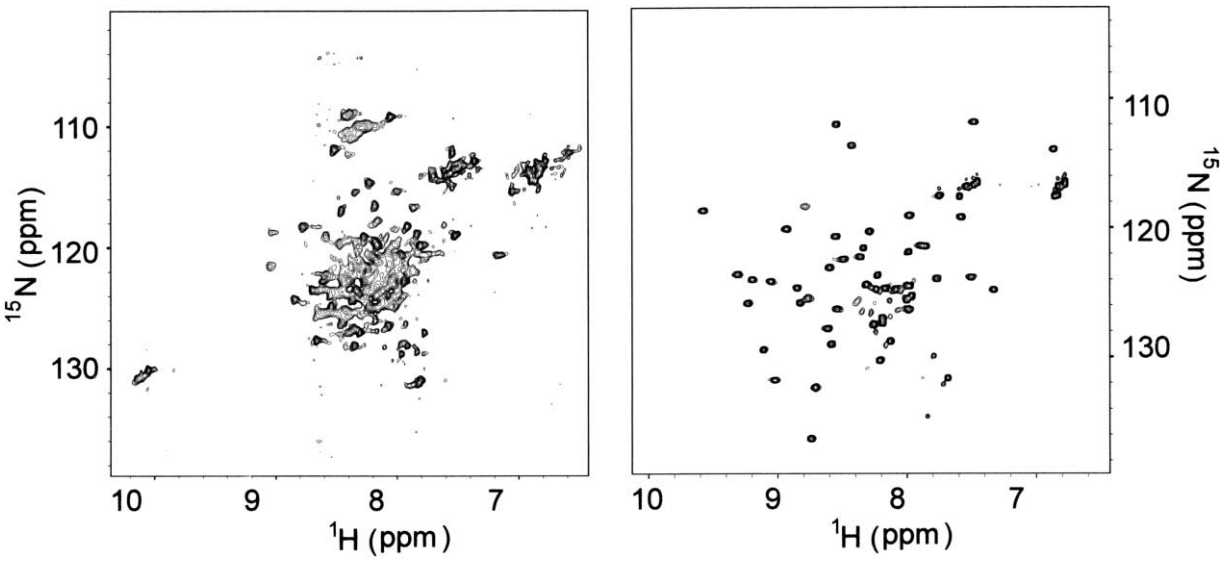

Figure 5. ^{15}N-HSQC Spectra of Unfolded and Folded Proteins

The left panel shows a ^{15}N-HSQC spectrum of a partially unstructured protein fragment of 80 amino acid residues. All signals cluster around a ^{1}H frequency of 8.3 ppm. Also, the signal dispersion in the ^{15}N dimension is limited. The broad unresolved signals in the middle of the spectrum indicate either aggregation in the sample or conformational heterogeneity on a ms-μs timescale (both cases are unfavorable for NMR studies). The signal at 10 ppm is not diagnostic for a folded protein but stems from the side chain amide group of a tryptophan residue. The right panel shows the spectrum of a folded, 55-residue-long construct of the IGFBP-5 protein. The peaks show a large signal dispersion in both dimensions.

corresponds to the overlap of slightly varying resonances from several fragments of different length. On the basis of these results, we cloned and expressed, in *E. coli*, a new fragment of the protein of 176 residues, which contained only the unchanged core of the protein. This fragment was not further degraded, even after several months. The spectrum of this protein fragment is shown in Figure 4C. Note the absence of the sharp resonances and the superior line width. NMR has revealed a stable folded core of the protein, which was then successfully subjected to the NMR and X-ray structure analysis.

The prominent signals from the small peptide fragments also provide an example for the examination of a sample's purity. Any small compounds, be they peptides or other impurities, will readily show in a one-dimensional spectrum.

Two-Dimensional NMR

Because of the greatly improved resolution of two-dimensional experiments, they are frequently used for screening and binding studies. The simplest and most powerful among them is the heteronuclear single-quantum coherence (HSQC) experiment. In an impressive large-scale approach, Yee et al. [15] have recently investigated more than 500 proteins from five different organisms, using ^{15}N-HSQC experiments to screen for those proteins amenable to NMR structure analysis. This spectrum is the first step in any structure elucidation, as it maps the backbone amide groups of a protein according to their proton and nitrogen frequencies. A whole set of three-dimensional spectra later used to assign the NMR signals to their respective amino acid residues is based on the HSQC experiment. For this kind of spectrum, ^{15}N-labeled protein samples are required. The HSQC shows one peak for every proton bound directly to a nitrogen atom and, thus, exactly one signal per residue in the protein (apart from proline, which is devoid of proton-bound nitrogen, and some additional side chain signals, which can easily be identified, appear).

The positions of the peaks are indicative of structured or disordered proteins in the same way as described above for the one-dimensional spectrum (Figure 5). In the spectrum of an unfolded protein, all signals cluster in a characteristic "blob" around a ^{1}H frequency of 8.3 ppm, with little signal dispersion in both dimensions. In the spectrum of a structured protein, the peaks show large signal dispersion. Thus, if the peaks are assigned their respective sequential positions in the polypeptide chain, disordered regions may be identified.

As the number of signals in the HSQC spectrum corresponds approximately to the number of residues in the protein under investigation, conformational heterogeneity can easily be detected by a surplus of peaks. To optimize sample conditions, pH titrations or titrations with cofactors or other molecules as well as variation of temperature may be performed while repeatedly recording HSQC spectra. This is feasible, since the NMR method is nondestructive and experiments may be repeated several times. It has, for example, been shown by NMR-observed titrations that low temperatures and neutral pHs stabilize the folded state of an SH3 domain of *Drosophila* drk, while high temperatures and low pHs tend to favor the unfolded state [26]. On the other hand low pH has also been reported to prevent aggregation, as observed by line width comparison [20].

For full NMR structure investigations, samples of 200–400 μl with a protein concentration of 0.5–1.0 mM are

required. This corresponds to about 10–15 mg/ml of the protein, which is the concentration usually used for crystallographic screening. Spectra can also be recorded in up to 1 mM Tris buffer. Note again that NMR does not destroy the sample, so it is possible to continue with crystallization attempts after NMR characterization.

Conclusion

NMR can serve as a powerful tool in structural proteomics, as a method for screening proteins for sample conditions that favor crystallization or that are candidates for the NMR solution structure determination.

Acknowledgments

This work was supported by a DFG grant (Program SFB 413).

Received: July 12, 2002
Revised: August 23, 2002
Accepted: September 17, 2002

References

1. Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C., and Szyperski, T. (2000). Protein NMR spectroscopy in structural genomics. Nat. Struct. Biol. *7*, 982–985.
2. Prestegard, J.H., Valafar, H., Glushka, J., and Tian, F. (2001). Nuclear magnetic resonance in the era of structural genomics. Biochemistry *40*, 8677–8685.
3. Sali, A. (1998). 100,000 protein structures for the biologist. Nat. Struct. Biol. *5*, 1029–1032.
4. Shuker, S.B., Hajduk, P.J., Meadows, R.P., and Fesik, S.W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. Science *274*, 1531–1534.
5. Stoll, R., Renner, C., Zweckstetter, M., Brüggert, M., Ambrosius, D., Palme, S., Engh, R.A., Golob, M., Breibach, I., Buettner, R., et al. (2001). The extracellular human melanoma inhibitory activity (MIA) protein adopts an SH3 domain-like fold. EMBO J. *20*, 340–349.
6. Renner, C., and Holak, T.A. (2001). NMR N-15 relaxation of the insulin-like growth factor (IGF)-binding domain of IGF binding protein-5 (IGFBP-5) determined free in solution and in complex with IGF-II. Eur. J. Biochem. *268*, 1058–1065.
7. Mühlhahn, P., Zweckstetter, M., Georgescu, J., Ciosto, C., Renner, C., Lanzendorfer, M., Lang, K., Ambrosius, D., Baier, M., Kurth, R., et al. (1998). Structure of interleukin 16 resembles a PDZ domain with an occluded peptide binding site. Nat. Struct. Biol. *5*, 682–686.
8. Wijesinha-Bettoni, R., Dobson, C.M., and Redfield, C. (2001). Comparison of the structural and dynamical properties of holo and apo bovine α-lactalbumin by NMR spectroscopy. J. Mol. Biol. *307*, 885–898.
9. Fielding, L. (2000). Determination of the association constants (Ka) from solution NMR data. Tetrahedron *56*, 6151–6170.
10. Zurdo, J., Guijarro, J.I., Jimenez, J.L., Saibil, H.R., and Dobson, C.M. (2001). Dependence on solution conditions of aggregation and amyloid formation by an SH3 domain. J. Mol. Biol. *311*, 325–340.
11. Canet, D., Last, A.M., Tito, P., Sunde, M., Spencer, A., Archer, D.B., Redfield, C., Robinson, C.V., and Dobson, C.M. (2002). Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. Nat. Struct. Biol. *9*, 308–315.
12. Diercks, T., Coles, M., and Kessler, H. (2001). Applications of NMR in drug discovery. Curr. Opin. Chem. Biol. *5*, 285–291.
13. Pellecchia, M., Sem, D.S., and Wüthrich, K. (2002). NMR in drug discovery. Nat. Rev. Drug Discov. *11*, 211–219.
14. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C.H., and Edwards, A.M. (2000). Structural proteomics: prospects for high throughput sample preparation. Prog. Biophys. Mol. Biol. *73*, 339–345.
15. Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G.M., Bhattacharyya, S., Gutierrez, P., et al. (2002). An NMR approach to structural proteomics. Proc. Natl. Acad. Sci. USA *99*, 1825–1830.
16. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., et al. (2000). Structural proteomics of an archaeon. Nat. Struct. Biol. *7*, 903–909.
17. Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (New York: Wiley).
18. Abragam A., (1961). Principles of Nuclear Magnetism, International Series of Monographs on Physics, No. 32 (New York: Oxford University Press).
19. Kalus, W., Zweckstetter, M., Renner, C., Sanchez, Y., Georgescu, J., Grol, M., Demuth, D., Schumacher, R., Dony, C., Lang, K., et al. (1998). Structure of the IGF-binding domain of the insulin-like growth factor-binding protein-5 (IGFBP-5): implications for IGF and IGF-I receptor interactions. EMBO J. *17*, 6558–6572.
20. Anglister, J., Grzesiek, S., Ren, H., Klee, C.B., and Bax, A. (1993). Isotope-edited multidimensional NMR of calcineurin B in the presence of the non-deuterated detergent CHAPS. J. Biomol. NMR *3*, 121–126.
21. Edwards, A., Arrowsmith, C.H., Christendat, D., Dharamsi, A., Friesen, J.D., Greenblatt, J.F., and Vedadi, M. (2000). Protein production: feeding the crystallographers and NMR spectroscopists. Nat. Struct. Biol. Suppl. *7*, 970–972.
22. Kalus, W., Baumgartner, R., Renner, C., Noegel, A., Chan, F.K.M., Winoto, A., and Holak, T.A. (1997). NMR structural characterization of the CDK inhibitor $p19^{INK4d}$. FEBS Lett. *401*, 127–132.
23. Baumgartner, R., Fernandez-Catalan, C., Winoto, A., Huber, R., Engh, R.A., and Holak, T.A. (1998). Structure of human cyclin-dependent kinase inhibitor $p19^{INK4d}$: comparison to known ankyrin-repeat-containing structures and implications for the dysfunction of tumor suppressor $p16^{INK4d}$. Structure *6*, 1279–1290.
24. Wishart, D.S., Sykes, B.D., and Richards, F.M. (1991). Relationship between nuclear magnetic shift and protein secondary structure. J. Mol. Biol. *222*, 311–333.
25. Gottwald, U., Brokamp, R., Karakesisoglou, I., Schleicher, M., and Noegel, A.A. (1996). Identification of a cyclase-associated protein (CAP) homologue in Dictyostelium dicoideum and characterization of its interaction with actin. Mol. Biol. Cell *7*, 261–272.
26. Zhang, O., and Forman-Kay, J.D. (1995). Structural characterization of folded and unfolded states of an SH3 domain in equilibrium in aqueous buffer. Biochemistry *34*, 6784–6794.